Fake reefs are sometimes *reefs* and sometimes not, but are always compositional

Hayley Ross, Najoung Kim & Kathryn Davidson*

Abstract. The semantics of adjective modification often begins with set intersection, such that ⟦yellow flower⟧ = ⟦yellow⟧∩⟦flower⟧. Thus a *yellow flower* is a *flower*. Such an account, however, runs into problems for adjectives like *fake* or *counterfeit*, which display a privative inference: a *fake gun* is not a *gun* and a *counterfeit dollar* is not a *dollar*. Moreover, recent work shows privativity cannot easily be encoded as a property of specific adjectives like *counterfeit*, since e.g. *counterfeit watch* robustly licenses the subsective inference of being a *watch* (Martin 2022). We gather judgments on nearly 800 adjective-noun bigrams (of which 180 are novel, i.e. zero corpus frequency), and show that privativity depends on the adjective, noun and context, and can be manipulated for the very same adjective-noun bigram by presenting it in different contexts. This poses a challenge for theories which fix privativity as a property of the adjective and always use the same method of composition (Partee 2010, del Pinal 2015). Moreover, we find no difference in participant behavior between novel adjective-noun bigrams and high frequency ones, suggesting that the process is nonetheless compositional and not the result of convention or memorized idiosyncrasy. Our results support compositional accounts like Martin (2022) (which modifies del Pinal 2015) and Guerrini (2024), which treat privativity as context-dependent.

Keywords. adjectives; nouns; compositionality; privativity; entailment; semantics

1. Introduction. A central concern for the study of meaning is how the meanings of complex expressions are composed from the meanings of their constituent parts. The fact that people understand completely novel phrases provides an argument that meaning must be governed by some kind of compositionality (Partee 2009). This paper, following in a growing tradition (Partee 2009, 2010, Szabó 2012, del Pinal 2015, i.a.), studies the dynamic interaction of meaning and context through the lens of (privative) adjective modification and how to account for it compositionally.

Historically, privativity has been defined as an adjective-specific phenomenon which negates the noun that the adjective combines with. A *fake gun* is said to be precisely *not a gun*. This property distinguishes privative adjectives from other types of adjectives, which typically yield an intersective or subsective inference. Canonical examples of privative adjectives include *fake, false, former, counterfeit, knock-off, mock,* and perhaps also *artificial* and *virtual* (Nayak et al. 2014).

(1) *Intersective inference*
This is a yellow flower.
∴ This is yellow.
∴ This is a flower.

(2) *Subsective inference*
This is a small elephant.
∴ This is an elephant.
∴̸ This is small.

(3) *Privative inference*
This is a fake gun.
∴ This is not a gun.

*Many thanks to research assistant Kate Bigley, and to Jesse Snedeker and all of the members of the Harvard Meaning & Modality lab for their helpful feedback. Special thanks to Josh Martin, whose dissertation on this topic inspired this project. This work was supported by an MBB Graduate Student Research Award from Harvard's Mind, Brain and Behavior Initiative. Authors: Hayley Ross, Harvard University (hayleyross@g.harvard.edu), Najoung Kim, Boston University (najoung@bu.edu) & Kathryn Davidson, Harvard University (kathryndavidson@fas.harvard.edu).

Privativity poses a challenge for compositionality, which requires the meaning of a complex expression to be derived solely from the meaning of its constituent parts (Szabó 2012): the set of *guns* combined with whatever function or set *fake* denotes. Modification of nouns by adjectives is classically treated as simple set intersection, shown in (4), as in textbooks like Heim & Kratzer (1998) and Coppock & Champollion (2023).[1] If a *fake gun* is not a *gun*, it is not clear how to derive the meaning of *fake gun* from the set of (real) *guns* through set operations such as subsection. This means that our compositional process cannot arrive at the meaning of *fake fire* simply by having *fake* yield some subset of the meaning of *fire*.

(4) $[\![\text{yellow flower}]\!] = [\![\text{yellow}]\!] \cap [\![\text{flower}]\!] = \{x : x \text{ is yellow}\} \cap \{x : x \text{ is a flower}\}$

In contrast to the conventional framing of privativity as an adjective-specific phenomenon, Martin (2022) shows that inference patterns for so-called privative adjectives vary depending on the noun used. For example, *counterfeit* may license a privative or subsective inference depending on the noun (and accompanying context).

(5) *Subsective inference*
This is a counterfeit watch.
∴ This is a watch.

(6) *Privative inference*
This is a counterfeit dollar.
∴ This is not a dollar.

This per-adjective and per-noun variation raises an additional question of whether these adjective-noun combinations and their inferences are computed (compositionally) on the fly, based on just the given adjective, noun and context, or whether there is an element of convention or past experience necessary to derive these varying inferences, in which case the inference would be stored (memorized). A significant body of processing work (Arnon & Snider 2010, Tremblay & Baayen 2010, Caldwell-Harris et al. 2012, O'Donnell 2015, i.a.) reveals plenty of cases where humans *don't* appear to compose meaning on the fly: chunks of various sizes from multi-morpheme words to entire idiomatic expressions, especially highly frequent words or expressions, can get stored as units and trigger priming effects in experimental studies, whether their meaning is idiomatic or fully compositional from their parts. If the effect of adjectives with privative inferences is stored rather than composed on the fly, then deriving the inference for infrequent adjective-noun bigrams with such adjectives, such as *fake scarf* or *fake reef*, might be difficult or result in widely varying results between people. The same might be true for intermediate, less memorization-heavy approaches such as learning (memorizing) the inferences for some high-frequency bigrams and then reasoning about novel bigrams by analogy where possible.[2] This paper explores the effect of experience (as measured by corpus frequency of the bigram) and context on adjective-noun combination and inferences, especially for novel (zero corpus frequency) adjective-noun bigrams.

We gather a large quantity of adjective-noun inference judgments for both high-frequency and novel / zero corpus frequency adjective-noun bigrams over three experiments and show that inferences depend not just on the adjective but also on the noun and the context. Further, we show that novel adjective-noun bigrams and their privativity inferences are handled as productively and

[1] To their credit, both textbooks note that gradable and/or non-intersective adjectives are not handled by this account.

[2] For example, a participant might reason that the novel bigram *counterfeit scarf* is a *scarf* by analogy to other clothing items and accessories such as *watch* or *handbag* which they have seen *counterfeit* occur with subsectively.

consistently by participants as those of high-frequency ones, despite the significant variation by adjective, noun, and context. Thus, any theory of adjective-noun combination must address the context-sensitivity of these inferences and predict the ability to generalize to novel combinations, for example using a compositional approach; it cannot be based on memorized idiomatic meanings. In Section 5, we discuss two compositional accounts of adjective-noun modification which satisfy these requirements and can handle the new data presented in this paper.

2. Choice of adjective-noun bigrams. We first establish a set of 798 bigrams to study, 23% of which are zero frequency in a large corpus, thus presumed novel for participants. The full set of bigrams, as well as results for all experiments in this paper, are available on GitHub.[3]

2.1. SELECTION BY CORPUS FREQUENCY. We consider 6 "privative" adjectives of interest: *fake, counterfeit, false, artificial, knockoff* and *former*. Since we established in the introduction that such adjectives need not always be privative, from here on out the phrase "privative adjective" will refer to these adjectives such as *fake* that have been discussed in prior literature as (typically) resulting in privative inferences. We select 6 intersective/subsective adjectives as "controls" which each have a similar frequency to one of the privative adjectives in a very large corpus (C4, ca. 130 trillion words; Raffel et al. 2020, Dodge et al. 2021) and which have relatively few selectional restrictions: *useful, tiny, illegal, homemade, unimportant and multicolored*. Frequencies are shown in Table 1. We choose *multicolored* as a low-frequency example of a (standardly intersective) colour adjective, *illegal* since it has negative valency while typically being subsective, and *homemade* since it targets the manner of manufacture, similar to *counterfeit* and *artificial*, without being obviously privative.

Adjective	Tokens	Adjective	Tokens
former	15.8M	useful	13.6M
false	4.6M	tiny	5.8M
artificial	3.9M	illegal	4.5M
fake	3.1M	homemade	2.2M
counterfeit	450K	unimportant	170K
knockoff	57K	multicolored	93K

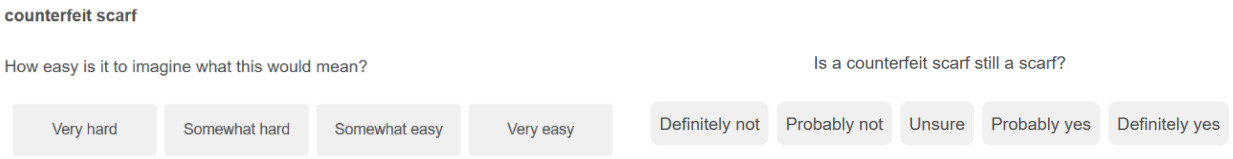
Table 1: Adjective frequencies in the C4 Corpus (130T words)

We algorithmically select 43 nouns from 300 nouns which commonly occur with a wide range of adjectives (Pavlick & Callison-Burch 2016a), plus the 36 nouns used in Martin (2022), with the goal of generating a high number of zero-frequency bigrams. We then manually select an additional 59 nouns which are semantically similar to these 43 nouns, for a total of 102 nouns. We cross these 102 nouns with the 12 adjectives for a total of 1224 bigrams to use in subsequent experiments. We determine (relative) bigram frequency by counting the frequency of all bigrams involved in this process in C4, for a total of 3979 bigrams (358 unique nouns $\times$ 12 adjectives, plus experiment fillers), since calculating the frequency over every possible corpus bigram would be prohibitively expensive. Thus, terms like "high frequency bigram" or "top quartile bigram" in this paper should be interpreted in relative rather than absolute terms.

[3] https://github.com/rossh2/artificial-intelligence

2.2. EXPERIMENT 1. Experiment 1 filters out clearly nonsensical combinations resulting from the blind crossing of adjectives and nouns. Combinations like *counterfeit accusation* or *multicolored effort* must be excluded before we can reasonably ask questions like "Is a counterfeit N still an N?" Participants are presented with a bigram and asked "How easy is it to imagine what this would mean?", as shown in Figure 1a.[4] 144 native American English speakers[5] were recruited on Prolific (of which 7 were excluded due to failed attention checks and/or not meeting the criteria for native English speaker); the study was implemented in Qualtrics. Participants were paid pro rata at $12/hour; the experiment took 4 minutes on average. Each participant saw 14 questions (12 target bigrams, 2 fillers), for 3 ratings per bigram in total.

We categorize bigrams whose ratings were majority "very hard" or "somewhat hard" as nonsensical, and exclude them from subsequent experiments. This leaves 798 bigrams, of which 23% (180) are zero frequency in C4, i.e. almost certainly novel to new participants, and another 21% (170) are low frequency (bottom quartile), so also quite possibly novel to participants.



(a) Experiment 1 (b) Experiment 2 on PCIbex

Figure 1: Screenshots of questions in Experiments 1 and 2

3. Experiment 2: *Is an A N an N?*.

3.1. METHOD. Experiment 2 asks participants *Is an A N still an N?* for each of the 798 adjective-noun bigrams left after filtering in Experiment 1. An example question is shown in Figure 1b. We choose to use the same design as Martin (2022), with the slight modification of adding *still* to make the question more natural.[6] This differs from previous privativity studies (Pustejovsky 2013, Pavlick & Callison-Burch 2016a,b) which ask participants to rate these inferences given a particular sentence context drawn from a corpus. For example, Pavlick & Callison-Burch (2016b) ask whether (7-a) entails (7-b). While more realistic than out-of-the-blue judgments, this also creates a much noisier picture, as demonstrated by this example: participants rate (7-a) as contradicting (7-b), but the verb *denied* and the world knowledge of pharmacists selling medicine also seem to be driving part of this inference. It actually remains unclear whether the counterfeit medicine that the pharmacists were selling qualifies as medicine in this scenario.

[4]Previous work studying novel adjective-noun combinations (Vecchi et al. 2017) uses a more complex pairwise ranking approach to precisely measure semantic deviance, but we only need to filter out obviously nonsensical bigrams.

[5]We recruit people on Prolific who self-report English as their first and primary language and are located in the United States. We further ask them at the end of the study whether they learned English before the age of 5 and whether they speak American English as opposed to another dialect of English (if not, they are paid but excluded). This implementation of "native speaker" is merely intended as a practical way to expect shared language experiences among our participant sample (Cheng et al. 2021).

[6]Using *still* seems to help foreground the idea that adding the adjective might change noun membership: a *fake scarf* or *unimportant sign* might *not* be a scarf or sign, or conversely might be a scarf *despite* being fake.

(7) a. Pharmacists in Algodones denied selling counterfeit medicine in their stores.
b. Pharmacists in Algodones denied selling medicine in their stores.

This experiment aims to show, for a wider range of adjectives and nouns than Martin 2022, that privativity varies depending on the noun, and investigates whether participants behave differently for high and zero frequency (assumed to be novel) bigrams. We will study non-out-the-blue judgments (in a more controlled setting than Pavlick & Callison-Burch) later, in Experiment 3.

We ran this experiment in two parts. For the first 305 bigrams, 510 native American English speakers were recruited on Prolific (of which 15 were excluded due to failed attention checks and/or not meeting the criteria for native English speaker); the study was implemented in PCIbex (Zehr & Schwarz 2018). For the next 498 bigrams, 756 native American English speakers were recruited on Prolific (of which 24 were excluded); this study was implemented on Qualtrics. Participants were paid pro rata at $12/hour and the experiment took 3 minutes on average. Each participant saw 12 questions (4 typically-intersective adjectives, 4 typically-privative adjectives, 4 fillers), for a total of approx. 12 ratings/bigram.[7] Since some bigrams which may not make sense to everyone likely remain after Experiment 1, we explicitly alert participants in Experiment 2 to this possibility and instruct them to use the "Unsure" rating if a combination does not make sense to them.

3.2. RESULTS. Mean bigram ratings are shown in Figure 2 (organized by adjective, each dot represents the mean rating for one adjective-noun bigram), and individual ratings by participants for a selection of bigrams are shown in Figure 3. We find that each so-called "privative" adjective in fact yields graded variation from privative to subsective depending on the noun, with ratings spanning all the way from 1 ("Definitely not [an N]") to 5 ("Definitely yes [an N]"). In Figure 3, we can also see that intermediate means are often associated with high variance rather than participants agreeing on "Unsure". We also find that "subsective" adjectives are usually subsective ("Probably yes" or "Definitely yes"), warranting the name, but are nonetheless not so clearly subsective with certain nouns (e.g. *homemade cat* with $\mu = 2.6$, *illegal currency* with $\mu = 2.83$).

Secondly, we find no effect of bigram frequency on rating variance. A linear regression in R (R Core Team 2023) shows that bigram frequency correlates poorly with the variance in the ratings (typically subsective: $R^2 = 0.003$, typically privative: $R^2 = 0.010$, both $p > 0.05$). Instead, participants agree to a similar degree on the meaning and inferences for high-frequency and zero frequency (novel) adjective-noun bigrams. Some high frequency bigrams such as *artificial tree* or *former house* show high variance in ratings ($\mu = 3.50, \sigma^2 = 1.83$ and $\mu = 3.63, \sigma^2 = 1.76$ respectively), suggesting that these bigrams do not have a conventionalized meaning or inference when presented out of the blue. Moreover, some zero frequency bigrams like *knockoff image* and *counterfeit scarf* have quite low variance ($\mu = 4.90, \sigma^2 = 0.10$ and $\mu = 4.80, \sigma^2 = 0.18$), suggesting that participants compose even novel bigrams systematically.

3.3. DISCUSSION. The results from Experiment 2 lend further weight to previous work illustrating that no adjective is unequivocally privative, but rather that privativity depends on the combination of adjective and noun (Martin 2022). We further see no correlation between rating variance and bigram frequency. We conclude that high frequency need not lead to a fixed conception of bi-

[7] Due to issues with PCIbex, the first part of experiment did not yield an even number of ratings per bigram. In the analysis of this experiment, we randomly sample and cap the number of ratings at (10-)12 ratings/item.

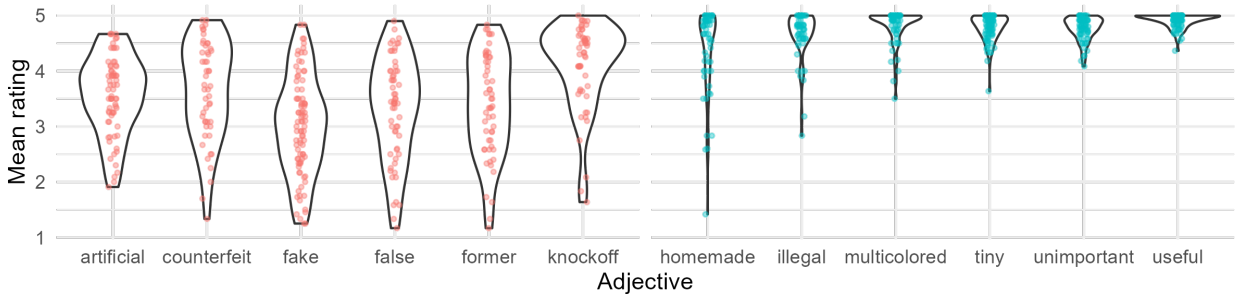


Figure 2: Mean ratings for "Is an AN an N?" for each bigram by adjective in Experiment 2, where 1 is most privative ("Definitely not") and 5 is most subsective ("Definitely yes").

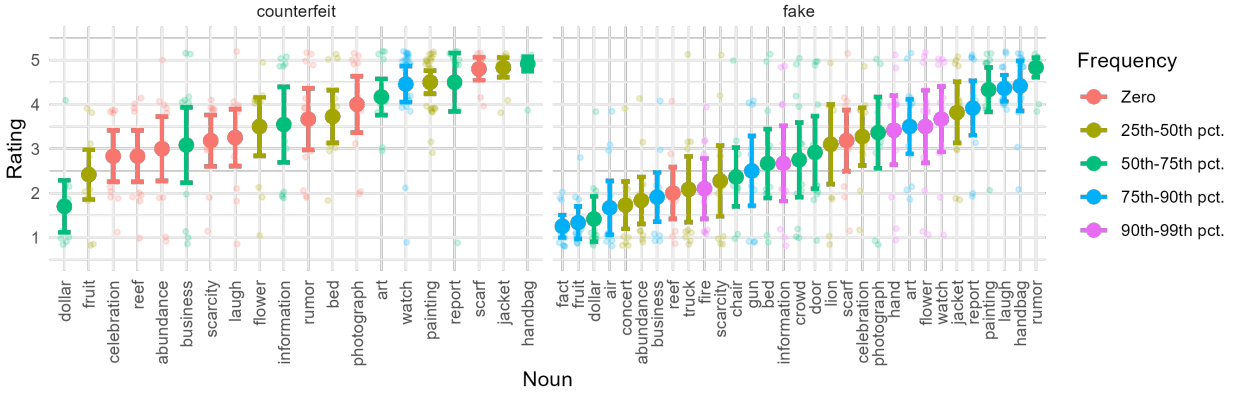
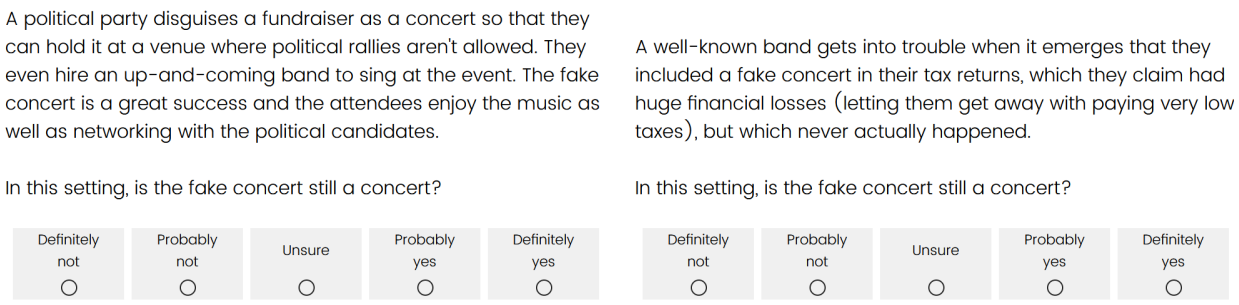


Figure 3: Participant ratings for a selection of bigrams involving *fake* and *counterfeit* in Experiment 2, where 1 is most privative and 5 is most subsective. • shows the mean with SE for each bigram.

gram meaning or inference, and that previous exposure to a (potentially) privative adjective-noun pair is not required to draw this inference, even for adjectives with relatively broad meanings like *fake*. Instead, we suspect that high variance may be due in part to participants imagining different contexts for the bigrams (which were presented out of the blue in Experiment 2), such that e.g. *fake* might target different aspects of the noun's properties or different properties might be relevant for noun-hood in that context. For example, a *fake crowd* might qualify as a *crowd* if it is made up of paid actors, but less so if it is just painted dummies on a movie set.

4. Experiment 3: Context.

4.1. METHOD. For 28 adjective-noun bigrams from Experiment 2, we construct two contexts each intended to bias the reader towards a subsective or privative inference respectively. Two example contexts for *fake concert* are shown in Figure 4. We targeted 6 pairs of adjective-noun bigrams from Experiment 2 with intermediate mean ratings and high variance, such that one bigram is zero/low frequency and the other is high frequency; we will use these pairs to investigate any effect of frequency. We then selected an additional 16 bigrams with intermediate mean ratings and high variance for which we were able to write convincing example contexts. We select



(a) Subsective-biased context (b) Privative-biased context

Figure 4: Screenshots of contexts for *fake concert* in Experiment 3.

these bigrams which are neither at ceiling or floor precisely because we suspect that there may be more than one context in participants' minds, and thus the specified contexts might split apart these middle-of-the-scale ratings into high (subsective) or low (privative) ratings respectively, explaining (some of) the variance in Experiment 2. Further, we select pairs of high and zero/low frequency bigrams because it is possible that high frequency bigrams might come with more conventionalized contexts and/or more fixed meanings and inferences in general, and thus might resist manipulation by provided contexts. Conversely, zero/low frequency bigrams might be particularly easy to manipulate, since they lack any preconceived "default" context.

For the first 12 bigrams, 40 native American English speakers were recruited on Prolific (of which 1 excluded due to failed attention checks); for the second set of 18 bigrams (two bigrams were rerun), a further 40 native American English speakers were recruited on Prolific (of which 2 were excluded for not meeting the native speaker criteria). Both studies were implemented in Qualtrics. Participants were paid pro rata at $12/hour; the experiment took 8 minutes on average. In the first instance of the experiment, each participant saw 12 items as shown in Figure 4; in the second instance, each participant saw 18 items (3 or 6 intersective-biased, 3 or 6 privative-biased, 6 fillers), yielding 10 ratings/item.

4.2. RESULTS. We find that across the board, writing biased contexts does indeed shift participants ratings in the intended direction, as shown in Figure 5, as well as reducing the variance. The one exception is *counterfeit dollar*, which refuses to be influenced by context at all. This can be explained simply by its meaning: *dollars* depend so heavily on having an authentic method of manufacture that any way in which they can be counterfeited, i.e. in which their method of manufacture is non-conventional, robs them of being a dollar. We fit an ordinal mixed effects model in R (R Core Team 2023, Christensen 2022) and find statistically significant effects for both the subsective-biased and privative-biased contexts compared to having no context ($p < 0.05$ for both; a subsective-biased context makes a high (subsective) rating 4x more likely while a privative biased context makes a high (subsective) rating only ¼x as likely). As in Experiment 1, we find no effect of frequency in this experiment: high-frequency bigrams do not have more fixed inferences and are not more resistant to being manipulated by context.

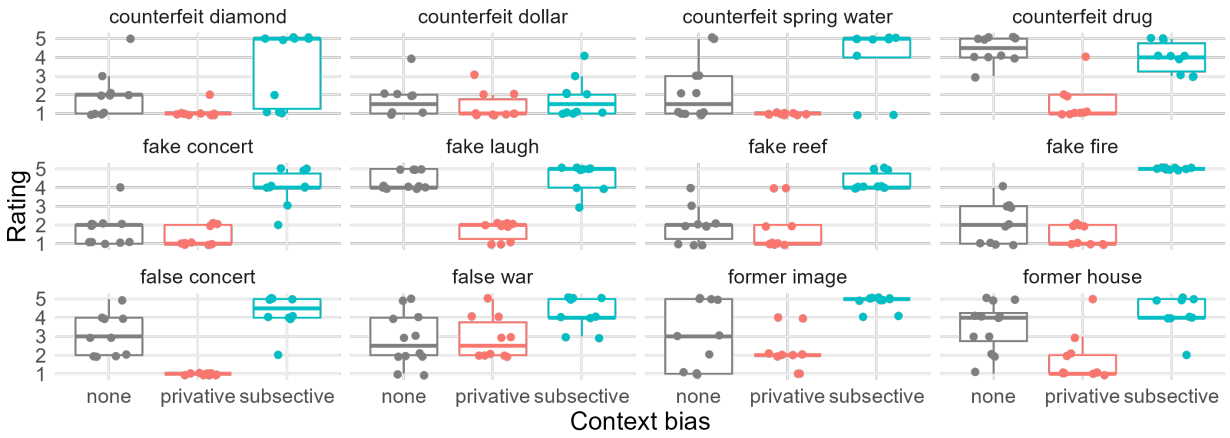


Figure 5: Ratings for "In this setting, is an AN still an N?" for 12 of 28 bigrams in Experiment 3, where 1 ("Definitely not") is most privative and 5 ("Definitely yes") is most subsective. The ratings from Experiment 2 are shown in gray. The first and third columns show zero or low frequency bigrams; the second and fourth columns show corresponding high frequency bigrams.

4.3. DISCUSSION. The results from Experiment 3 show that the subsective/privative inferences drawn from adjective-noun combination are indeed context-dependent: providing different contexts can cause participants to draw quite subsective or privative inferences for the very same adjective-noun bigram. This is possible whether the bigram is high-frequency (thus potentially coming with a "default" or conventionalized context of use) or novel. Thus it is likely that the participants' imagined contexts explains some of the variance in Experiment 2, which presented the bigrams out of the blue. In other words, the meaning of adjective-noun bigrams and their privative inferences cannot be explained by memorization of a single (conventionalized) meaning or inference, since inferences must be computed productively on the fly based on the provided context (as well as world knowledge). Finally, the ability to manipulate the inferences of novel bigrams such as *false concert* supports a compositional account for the meaning of the bigram (where context is included as part of the composition and inference-drawing process), as in Experiment 2.

5. Impact on theoretical accounts. Our experiments showcase the wide variation in privative inferences among so-called privative adjectives: first, adjectives can license either a subsective or privative inference depending on the noun (and context), and second, the same adjective-noun bigram can license either inference depending on context. These results support a compositional, context-dependent account of adjective-noun modification rather than an approach based on prior experience or convention which memorizes the meanings and/or inferences for previously encountered bigrams. Further, these results pose challenges for theories which treat privativity as a property of only the adjective, such as Partee (2010), del Pinal (2015) and Guerrini (2024).

5.1. PARTEE'S NON-VACUITY PRINCIPLE. Partee's classic account of privative adjectives (Kamp & Partee 1995, Partee 2007, 2009, 2010) posits that all seemingly privative adjectives in fact compose subsectively with the noun. Unlike regular subsective adjectives, however, $[\![\text{fake gun}]\!] = \varnothing$ initially, since *gun* includes only real guns and *fake* is privative (by definition). The *Non-Vacuity*

Principle then coerces *gun* to expand to include both real and fake guns, so that *fake* can now act subsectively over this new expanded set. Since this account stipulates that *fake* is always privative with respect to the original noun, it is unclear how it accounts for our experimental data that e.g. a *fake watch* is frequently a *watch*. While the intuitions behind Partee's account seem to be on the right track and have inspired several subsequent analyses, her specific implementation and the Non-Vacuity Principle do not capture the full extent of our data, in addition to existing criticisms of her implementation (Martin 2022, del Pinal 2015, 2018, Guerrini 2024).

5.2. QUALIA-BASED ACCOUNTS. Del Pinal (2015, 2018) argues that we can have a compositional, truth-conditional account and still explain the behaviour of adjectives like *real, typical* and *fake* in a principled manner by moving to a two-dimensional semantics. Del Pinal's Dual Content Semantics enriches lexical entries to have an extensional component (E-structure), which corresponds to the traditional set extension, plus a conceptual component (C-structure). This C-structure essentially captures the concept behind the noun or adjective, leaning on the large body of literature on concepts in psychology to do so. Del Pinal implements it using qualia.

Adjectives like *fake* draw on the C-structure of nouns like *gun* to build the new E-structure for *fake gun*. Specifically, *fake* modifies *gun* to yield the semantics in (8) for *fake gun*: firstly, in the E-structure, *fake guns* are not in the extension of guns. Secondly, *fake guns* do not have the origins of guns (the agentive property), instead, they were made to appear as if they were guns. In the C-structure, this is also the new agentive property, and means that they have the appearance of guns (the same formal properties) and do not have the same purpose (telic property).

(8) $[\![\text{fake gun}]\!]$ = (del Pinal 2015; p.21)

E-structure: $\lambda x.\, \neg Q_E([\![\text{gun}]\!])(x) \wedge \neg Q_A([\![\text{gun}]\!])(x)$
$\wedge\, \exists e_2\, [\text{making}(e_2) \wedge \textsc{goal}(e_2, Q_F([\![\text{gun}]\!])(x))]$

C-structure:

CONSTITUTIVE:	$Q_C([\![\text{gun}]\!]) = \lambda x.\, \text{parts-gun}(x)$
FORMAL:	$Q_F([\![\text{gun}]\!]) = \lambda x.\, \text{perceptual-gun}(x)$
TELIC:	$\neg Q_T([\![\text{gun}]\!]) = \lambda x.\, \neg\textsc{Gen}\, e[\text{shooting}(e) \wedge \text{instrument}(e, x)]$
AGENTIVE:	$\lambda x.\, \exists e_2\, [\text{making}(e_2) \wedge \textsc{goal}(e_2, Q_F([\![\text{gun}]\!])(x))]$

While del Pinal's theory lays out in much more detail than Partee how the composition works, del Pinal (2015) still stipulates, like Partee, that *fake guns* are not *guns* by fixing in the E-structure that a *fake gun* is not in the extension of *gun*: $\neg Q_E([\![\text{gun}]\!])(x)$. In subsequent work, del Pinal (2018) admits that this part of the E-structure is questionable for *counterfeit* and *artificial* and that it is an empirical question whether this should be included or not. Empirically, both we and Martin (2022) find that it should not be included for any "privative" adjective.

Martin (2022) adjusts del Pinal's definitions to remove privativity in the E-structure. Instead, privativity arises after composition of the adjective and noun, when set membership of the newly composed object is determined. If a targeted, negated quale (primarily TELIC for *fake*) is "central" to the meaning of the noun, i.e. part of the E-structure, then this results in privativity. Which qualia (analogous to "typical properties" of the noun) are incorporated into the E-structure is context-dependent, as it is for any use of a noun. This allows the modified account to capture the variation in our experiments by noun and by context.

5.3. SIMILARITY-BASED ACCOUNTS. Guerrini (2024) presents a similarity-based account of *fake*, which builds on the intuitions for *fake* and *counterfeit* in del Pinal (2015), but implements the lexical entry using using the one-dimensional compositional semantics for *seems like* from Guerrini (2022) instead of using a two-dimensional semantics with qualia.

(9) $[\![\text{fake}]\!] = \lambda P \lambda x.\, \textsc{Intended}(x, \textsc{seem-like}(x, P)) \wedge \neg P(x)$ (Guerrini 2024; p.190)

(10) $\textsc{Intended}(x, P) = \exists e.\, \textsc{Action}(e, x) \wedge \textsc{Goal}(e, P(x))$

Note that like the original definition of *fake* in del Pinal (2015), and in keeping with Partee, privativity is explicitly part of Guerrini's definition of *fake*. Unlike del Pinal, Guerrini is committed to this fact. He argues that the variability in privativity in the experiments in Martin (2022), which we extend in this paper, is in fact explained by syntactic ambiguity. Building on the account in Martin (2022) for the ambiguity between subsective and intersective readings of e.g. *good thief* (good at thieving, or a good person and a thief), Guerrini argues that when *fake watch* has a subsective reading, *fake* (privatively) modifies another covert, contextually supplied noun, such as *Rolex*: syntactically, we have [$_{\text{NP}}$ [$_{\text{AP}}$ *fake* ~~*Rolex*~~] *watch*]. Since this syntactic ambiguity is context-dependent, Guerrini (2022) in principle also accounts for the data presented in Experiments 2 and 3. One concern with this account is whether the contextually supplied material has to be a single noun over which *fake* acts privatively. Our data suggests this would be difficult for items such as *fake concert* in Figure 4, though there may be a multi-word phrase or concept that can satisfy the theory.

6. Conclusion. This paper presents experimental evidence on the variation in subsective vs. privative inferences both within adjectives and within adjective-noun bigrams, including for novel adjective-noun bigrams. We find that no adjective always yields privative inferences, lending further weight to Martin (2022). We further find that for most adjective-noun combinations, privativity depends on the context as well as the adjective and the noun. Our results show that any theory of adjective-noun combination must account for the context-sensitivity of these inferences and allow generalization to novel combinations (for example, by composition)—these inferences are not so unpredictable as to need to be memorized. Theories of privativity which are not compositional (e.g. basic set complementation) or which fix privativity as a property of individual adjectives (Partee 2010, del Pinal 2015) with only a single method of composition do not account for the full set of our data. Compositional accounts like Martin (2022)'s modification of del Pinal (2015, 2018) and Guerrini (2024) which predict that privativity is context-dependent, either by having privativity arise outside of the composition or by appealing to syntactic ambiguity, are able to account for the generalization and context-sensitivity found in our experiments. These compositional approaches aim to explain why participants are equally able to draw inferences for novel bigrams as for high-frequency ones, and what inferences we should expect given the effect of the context on available nouns or restrictions of noun denotations.

References

Arnon, Inbal & Neal Snider. 2010. More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language* 62(1). 67–82. https://www.sciencedirect.com/science/article/pii/S0749596X09000965.

Caldwell-Harris, Catherine, Jonathan Berant & Shimon Edelman. 2012. Measuring Mental En-

trenchment of Phrases with Perceptual Identification, Familiarity Ratings, and Corpus Frequency Statistics. In Dagmar Divjak & Stefan Th. Gries (eds.), *Frequency effects in language representation*, vol. 2, 165–194. Berlin: De Gruyter Mouton. https://www.degruyter.com/document/doi/10.1515/9783110274073.165/html.

Cheng, Lauretta S. P., Danielle Burgess, Natasha Vernooij, Cecilia Solís-Barroso, Ashley McDermott & Savithry Namboodiripad. 2021. The Problematic Concept of Native Speaker in Psycholinguistics: Replacing Vague and Harmful Terminology With Inclusive and Accurate Measures. *Frontiers in Psychology* 12. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.715843/full.

Christensen, Rune Haubo Bojesen. 2022. ordinal—Regression Models for Ordinal Data. Manuscript.

Coppock, Elizabeth & Lucas Champollion. 2023. Invitation to Formal Semantics. https://eecoppock.info/semantics-boot-camp.pdf.

Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell & Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. http://arxiv.org/abs/2104.08758.

Guerrini, Janek. 2022. 'Like a N' constructions: genericity in similarity. In Dean McHugh & Alexandra Mayn (eds.), *Proceedings of the ESSLLI 2022 Student Session*, Galway, Ireland. https://doi.org/10.21942/uva.20368104.

Guerrini, Janek. 2024. Keeping Fake Simple. *Journal of Semantics* 41(2). 175–210. https://doi.org/10.1093/jos/ffae010.

Heim, Irene & Angelika Kratzer. 1998. *Semantics in generative grammar* Blackwell textbooks in linguistics 13. Malden, Mass., USA: Blackwell.

Kamp, Hans & Barbara H. Partee. 1995. Prototype theory and compositionality. *Cognition* 57(2). 129–191.

Martin, Joshua. 2022. *Compositional Routes to (Non)Intersectivity*. Cambridge, MA: Harvard University dissertation. https://www.proquest.com/docview/2681380157/abstract/58A63B8C3E6548AEPQ/1.

Nayak, Neha, Mark Kowarsky, Gabor Angeli & Christopher D. Manning. 2014. A Dictionary of Nonsubsective Adjectives. Tech. Rep. CSTR 2014-04 Department of Computer Science, Stanford University. https://www-cs.stanford.edu/~angeli/papers/2014-tr-adjectives.pdf.

O'Donnell, Timothy J. 2015. *Productivity and Reuse in Language: A Theory of Linguistic Computation and Storage*. Cambridge, Massachusetts, London, England: The MIT Press. https://doi.org/10.7551/mitpress/9780262028844.001.0001.

Partee, Barbara H. 2007. Compositionality and coercion in semantics: The dynamics of adjective meaning. In Gerlof Bouma, Irene Krämer & Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 145–161. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Partee, Barbara H. 2009. Formal semantics, lexical semantics, and compositionality: The puzzle of privative adjectives. *Philologia* 7(1). 11–21. http://www.philologia.org.rs/index.php/ph/article/view/216.

Partee, Barbara H. 2010. Privative adjectives: Subsective plus coercion. In Thomas Zim-

mermann, Rainer Bauerle & Uwe Reyle (eds.), *Presuppositions and discourse: Essays offered to Hans Kamp*, 273–285. Leiden, The Netherlands: Brill. https://brill.com/downloadpdf/book/edcoll/9789004253162/B9789004253162-s011.pdf.

Pavlick, Ellie & Chris Callison-Burch. 2016a. Most "babies" are "little" and most "problems" are "huge": Compositional Entailment in Adjective-Nouns. In Katrin Erk & Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2164–2173. Berlin, Germany: Association for Computational Linguistics. https://aclanthology.org/P16-1204.

Pavlick, Ellie & Chris Callison-Burch. 2016b. So-Called Non-Subsective Adjectives. In Claire Gardent, Raffaella Bernardi & Ivan Titov (eds.), *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, 114–119. Berlin, Germany: Association for Computational Linguistics. https://aclanthology.org/S16-2014.

del Pinal, Guillermo. 2015. Dual Content Semantics, privative adjectives, and dynamic compositionality. *Semantics and Pragmatics* 8. 7:1–53. https://semprag.org/index.php/sp/article/view/sp.8.7.

del Pinal, Guillermo. 2018. Meaning, modulation, and context: a multidimensional semantics for truth-conditional pragmatics. *Linguistics and Philosophy* 41(2). 165–207. https://doi.org/10.1007/s10988-017-9221-z.

Pustejovsky, James. 2013. Inference Patterns with Intensional Adjectives. In Harry Bunt (ed.), *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, 85–89. Potsdam, Germany: Association for Computational Linguistics. https://aclanthology.org/W13-0509.

R Core Team. 2023. R: A Language and Environment for Statistical Computing. https://www.R-project.org/.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li & Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21(140). 1–67. http://jmlr.org/papers/v21/20-074.html.

Szabó, Zoltán Gendler. 2012. The case for compositionality. In Wolfram Hinzen, Edouard Machery & Markus Werning (eds.), *The Oxford Handbook of Compositionality*, 64–80. Oxford: OUP. https://doi.org/10.1093/oxfordhb/9780199541072.013.0003.

Tremblay, Antoine & Harald Baayen. 2010. Holistic Processing of Regular Four-word Sequences: A Behavioural and ERP Study of the Effects of Structure, Frequency, and Probability on Immediate Free Recall. *Perspectives on Formulaic Language: Acquisition and Communication* 151–173.

Vecchi, Eva M., Marco Marelli, Roberto Zamparelli & Marco Baroni. 2017. Spicy Adjectives and Nominal Donkeys: Capturing Semantic Deviance Using Compositionality in Distributional Spaces. *Cognitive Science* 41(1). 102–136. https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12330.

Zehr, Jeremy & Florian Schwarz. 2018. PennController for Internet Based Experiments (IBEX). https://doi.org/10.17605/OSF.IO/MD832.